Exploring the use of Paragraph-level Annotations for Sentiment Analysis of Financial Blogs

Paul Ferguson[1], Neil O'Hare[1], Michael Davy[2], Adam Bermingham[1], Scott Tattersall[3], Paraic Sheridan[2], Cathal Gurrin[1], and Alan F. Smeaton[1]

[1]CLARITY: Centre for Sensor Web Technologies, Dublin City University, Ireland
[2]National Centre for Language Technology, Dublin City University, Ireland
[3]Zignals, Dublin, Ireland

Abstract. In this paper we describe our work in the area of topic-based sentiment analysis in the domain of financial blogs. We explore the use of paragraph-level and document-level annotations, examining how additional information from paragraph-level annotations can be used to increase the accuracy of document-level sentiment classification. We acknowledge the additional effort required to provide these paragraph-level annotations, and so we compare these findings against an automatic means of generating topic-specific sub-documents.

Key words: Sentiment Analysis, Opinion Mining, Financial Blogs

1 Introduction

The use of sentiment analysis is becoming increasingly useful as a means of automatically analysing the vast amounts of information available on the Web, and is of keen interest to many, for reasons such as marketing and financial market prediction. Our own work in sentiment analysis has been in the financial domain, and although this in itself is not new – previous work [1, 6] has been applied to traditional news and financial media – our work explores the use of sentiment analysis on articles from financial blogs. The blogosphere is widely acknowledged as a source of subjective opinions, as recognised in the TREC Blog Track [12], which has been run since 2006. Due to the subjective nature of blogs, they have the advantage that their authors are more likely to express opinions and make predictions about future performance, compared with traditional news sources which generally focus on reporting news relating to current or past performance.

This work is part of a collaboration between Dublin City University (DCU) and an industrial partner working in online stock trading[1]. The aim is to automatically extract the subjective opinions uniquely found on blogs and track the changing sentiment from the blogosphere towards individual stocks and the market in general.

[1] Zignals: http://www.zignals.com

We follow a supervised machine learning approach to automatic sentiment classification, which uses user annotations to train and test a classifier. The main focus of this paper is to investigate the use of paragraph-level annotations to impart more concise labelled training data to a supervised learning module. These annotations are more detailed than the typical document-level annotations, and we wish to investigate the effect they can have on the accuracy of the document classifier. Due to the freeform nature of blogs, articles often discuss multiple topics, and although the paragraph annotations should be helpful in dealing with the problem of topic shift, annotations at the paragraph level can be more time consuming to generate. Therefore we also compare the results produced with the use of paragraph-level annotations against our previous work [11], which provides a means of automatically extracting topic-specific sub-documents.

The paper is organised as follows. Section 2 gives an overview of related work, followed by details of our financial blog corpus and a description of the different types of annotations that were carried out on this corpus in Section 3. Section 4 describes how we use our paragraph annotations in order to predict sentiment classification, following by an evaluation of these approaches in Section 5. Section 6 draws conclusions from our work.

2 Related Work

Our approach relies on text annotated for sentiment to train a classifier. In the absence of author annotations, as is often available for sentiment analysis in certain domains [15], we rely on annotators to provide sentiment judgements. In the literature, text has been annotated at a number of different levels of granularity including document- [10], sentence- [18] and phrase-level [21]. It is unclear which granularity provides an optimum training scenario, though it is usual to annotate, train and test at the same granularity. It is also unclear, if when charged with a document-level annotation task, a classifier can benefit from annotations at a more fine-grained level.

An example can be found in the information retrieval literature where Santos [17] compared the performance of a sentence-level classification approach to a statistical document-based approach for document-level subjectivity ranking. The sentence classifications were provided by the OpinionFinder tool [20], which uses a combination of machine learning, lexicon and NLP techniques to classify sentences as subjective or objective, as well as providing polarity classfications. Santos found that while the sentence classification method provided an increase in performance above baseline, this technique did not outperform a simpler statistical approach. It should be noted, however, that subjectivity classification is a different task to sentiment polarity classification and that our challenge is to classify documents explicitly rather than produce a ranked list of documents.

Our work is in the financial blog domain, where sentiment analysis work has concerned identifying affect in news [3], identifying positive or negative news [1] and predictive analysis [7]. Our approach is unique in this domain as we endeavour to mine subjective user-generated content authored by bloggers to

provide a sentiment indicator for topics of interest. In our previous work [11], we hypothesised that the nature of the topic drift in this domain is such that providing the classifier with sub-sections of documents, relevant to the topics in question, would benefit our sentiment analysis, and showing an increase in classifier accuracy using such sub-documents.

Automatically detecting sections of the document pertinent to the topic is essentially ascribing a logical, topic-based structure to the document. Our corpus also has an inherent logical structure provided by the author in the form of paragraphs and we have annotated paragraphs (in addition to documents) for sentiment, providing annotations that we would intuitively expect to be more precise, as the sentiment is explicitly associated with shorter sections of text.

On the other hand, machine learning text categorisation relies on statistical approaches to model the overlying tone and style in the document. Reducing the proportion of document text provided during the training phase inherently runs the risk of omitting potentially discriminative text from the training documents. Indeed, the author of a blog is putting forth their thoughts in the form of a document, the sentiment of which may be something more subtle or high-level than simply the sum of the sentiment expressed in individual sentences or paragraphs. Thus we set out to answer the following research questions:

- How are a document's paragraph-level annotations best used to form a single document-level classification?
- How does this compare to traditional classification, where annotation, training and classifcation is at the document-level?

3 Financial Blog Corpus

The corpus that we use is composed of blog articles from 232 financial blog sourcs [11], crawled on two separate occasions: for three weeks in February 2009 and five weeks from May to June 2009, resulting in a collection of 7,757 documents. From this set of documents we annotated a subset for sentiment towards given topics. Our ultimate aim is to automatically determine the polarity of sentiment towards the set of financial stocks contained within the Standard & Poor's Index (S&P 500), which therefore are our topics of interest in this work.

3.1 Annotations

The use of human annotation is typically part of the process necessary in a supervised learning approach to sentiment analysis. The annotator generally annotates a document with the polarity of sentiment that it contains with regard to a pre-defined topic. As part of our work, in addition to annotating each of the documents, we have also annotated each of the paragraphs within the documents with their sentiment towards the same topics. Since these paragraph-level annotations are more concise than document-level annotations, and as they also give more information to train the learning module with, it is hoped that they can be of benefit in increasing the classifier's accuracy.

Due to the multi-topical nature of blogs in general and of the articles in our corpus (discussed in [11]), we cannot assume that a single article discusses only one topic. We therefore need to classify sentiment towards different topics from a single document. This produces a set of document-topic (or doc-topic) pairs which capture the sentiment in a particular document towards a specific topic.

We generated a list of 1526 unique document-topic pairs, from which we we carried out a total of 1691 annotations (some document-topic pairs were annotated twice to allow for inter-annotator agreement analysis). Each of these doc-topic pairs each annotator annotated for sentiment, using a five-point scale from *Very Negative* to *Very Positive*: *Very Negative, Negative, Neutral, Positive, Very Positive*, in addition to the labels *mixed*, which indicates a mixture of positive and negative sentiment, and *not relevant*. This annotation was carried out for each document, and for every paragraph within each document. Although more time-consuming to annotate, the paragraph annotations should be more accurate, particularly considering the issue of topic drift. In [11] we examined the variation in inter-annotation agreement, as the granularity of labelling was changed: a Kappa score of 0.712 was achieved for a 3-point labelling scheme, with a Kappa of 1.0 for binary annotations, while agreement for more fine-grained labelling scheme (i.e. the five point scale above) was quite low at 0.59.

4 Sentiment Classification Using Paragraph-level Annotations

In this section we describe how the typical document-level annotations were used to train and test a classifier, followed by a description of how this process was modified to make use of the paragraph-level annotations.

There are two distinct approaches presented in the literature to allow automatic sentiment polarity classification. The first uses a domain independent lexical resource to classify text [19, 2, 3], while the other builds up domain dependent models using machine learning techniques [13, 8, 5]. We focus on the use of a machine learning approach; we use a multinomial naïve Bayes (MNB) classifier, since it outperformed an SVM classifier in our previous work [11].

The classification task attempts to model a function $f : X \mapsto Y$ which maps from doc-topic pairs (X) to a set of predefined categories (Y). We explore two classification tasks:

- *Binary classification*, which predicts whether an article is either positive or negative to a given topic ($Y \in \{positive, negative\}$).
- *3-Point classification*, which is a finer level of classification granularity. In this case we include neutral documents ($Y \in \{positive, negative, neutral\}$).

When using the document-level annotations it is straight-forward to run training and testing phases on the data: however with the use of the paragraphs it can be somewhat more complex, as ultimately we want the predictions to be made at the document level, i.e. if we have a new document we want to be able to classify the document's overall sentiment with regard to a topic.

4.1 Topic-Based Text Extraction

Since blog articles often contain discussion of multiple topics, it is useful to extract those segments from the documents that are most relevant to the topic of interest. We previously explored using topic-based text extraction to enable sentiment analysis to be carried out at a sub-document level, ensuring that we restrict our analysis to the portions of a document relevant to a specified topic [11]. Our text extraction algorithms take a topic (a text string containing one or more terms) and extract sub-segments of the document that occur adjacent to any of the topic terms. We implemented three approaches: word-based, sentence-based, and paragraph-based methods, which extract word-, sentence- or paragraph- windows of size n either side of the target topic. Previously, we applied such text extraction techniques to documents (i.e. sub-segments were extracted from documents). Those approaches are used as a baseline for comparison in this work; when training using paragraphs we also explore the utility of using word-based text extraction approaches to extract the most relevant sub-segments of paragraphs for training and testing.

4.2 Paragraph Training / Document Testing

The first approach to using use paragraph-level annotations to train a classifier, while producing document-level classifications at the testing phase, is the simple approach of providing the paragraph text and annotations as inputs during training, which builds the classifier. Then during the testing phase we use the document text and annotations to test the classifier that was built using paragraph data. The motivation for this is that the more concise, but higher volume, training data provided by the paragraph annotation will lead to a better classification model, which document classification can then benefit from.

4.3 Paragraph Aggregation

The next approach is to carry out both training and testing using the paragraph-level data, but then to aggregate the results for paragraphs relating to the same topic from each document . This can be seen as a data fusion step in which we take information from multiple sources (in this case classifications for different paragraphs) and from these we generate a single classification for a document. We experiment with the use of two different aggregation schemes:

- *Sum of Predictions:* summing of prediction scores for each sentiment polarity class (from the MNB classifier). For each classification the MNB classifier gives its probability of belonging to each class (e.g. positive, neutral, negative). As each document consists of multiple paragraphs, if we sum the probability scores for each class across all paragraphs (for each document) then we can gain an overall document prediction score for each class – from this we then choose the class with the highest score.

– *Majority Voting:* taking a majority vote of the predictions, based on all predictions generated for the paragraphs within a document. In the case of the same number of paragraphs having the same number of predictions/votes the ties are broken using the Sum of Predictions approach above.

5 Evaluation

We evaluated our sentiment analysis approaches using the financial blog corpus described in Section 3 (and in more detail in [11]). Examples not having the labels Y (see Section 4 above) are discarded, while those examples that were labelled inconsistently by more than one annotator are also discarded. This leaves a total of 687 labelled document-topic pairs and 1629 paragraph-topic pairs for binary classification, while for 3-Point classification we have 917 labelled document-topic pairs and 2402 paragraph-topic pairs. From this we use ten-fold cross validation for each of the experiments using classication accuracy as the performance metric (with the results averaged over the ten folds).

As a pre-processing step, the dataset was firstly tokenised on whitespace, digits and punctuation characters. Following this, we removed stopwords using the list from the RCV1 [9] corpus, stemmed all tokens using the Porter stemming algorithm [16], and transformed all tokens to lowercase. We used the bag-of-words representation to construct feature vectors for each document and sub-document. A binary weighting scheme was employed since we found that, similar to previous research, binary weighting (i.e. terms are labelled as being present/not present, ignoring frequency information) delivered better performance than alternative weighting schemes [4, 14].

Firstly, we carry out experiments using the paragraph-level data to classify paragraphs, then concentrate on using the same paragraph-level annotation data to train classifiers which are used to classify sentiment at the document-level.

5.1 Paragraph Classification

In Table 1, we present paragraph classification results using the full paragraph text, while the additional information shows the performance if we employ the word-based text extraction approaches (which were found to work best on the full document text when classifying at the document level [11]).

From this it is clear that although we can achieve a relatively high classification accuracy (79% for binary classification), the results of the word-based text extraction show that this afford any extra improvement: in fact, it degrades performance. This is contrary our previous results using topic-based text extraction with full documents [11]. However, this possibly illustrates that the gains of using text extraction are only to be seen when segmenting a full document to create a more concise sub-document, and when working with an already concise paragraph then there are no additional gains to be achieved.

	Words	
n	**Binary**	**3-Point**
20	77.857	61.1597
30	76.8306	60.2696
40	77.1784	60.6579
50	78.5444	62.2083
60	78.375	61.9885
Full Paragraph	**79.0196**	62.9231

Table 1. Binary and 3-Point paragraph classification results using paragraph annotation data for training and testing, with and without word based text extraction. The 'n' column indicates the size of the word window used.

5.2 Using Paragraph Annotations for Document Classification

Although it is interesting to analyse the performance of the paragraph classification, we are ultimately interested in examining the author's overall option towards a topic, based on the full document. Therefore, if we wish to use the paragraph-level information we will have to provide a means of using paragraph annotations to train our classifier, then classify at the document-level at the classification testing phase.

Firstly, in Table 2, we present the most straight-forward approach of using the paragraph-level data in the training phase and then testing on the document-level data. We can see that this approach degrades the performance of the baseline "Document Training" – which is achieved using the full document-level annotation data to do both the training and testing.

	Binary	**3-Point**
Document Training	**69.5447**	54.454
Paragraph Training	63.4213	49.5139

Table 2. Document classification using the full document annotations vs. paragraph annotations for training.

Next we investigate the use of text extraction applied to the testing documents, in the same way is this was done in our previous work [11], while still using the paragraph annotations for training. We concentrate on word- and paragraph- based text extraction, which previously gave the optimal performance. This provides a topic-specific sub-document based on a window of words around the topic term. In Table 3 we present two (baseline) results under the "Document Training" heading which both use the documents as training and testing: the first result is as shown previously (using the full document) and the second result shows our optimal result using word based text extraction to create a topic-specfic dsub-aragraph-level data as training and using a segmented

version of the document at testing – text extraction using word windows and paragraph windows around the topic words. We present the optimal range of word window values (as found in previous work), i.e. paragraph (n = 0, i.e. only paragraphs containing the topic) and word (n = 20, 30. 40, 50). We can see that although there is a marginal improvement gained with the binary classifier using paragraph-based text extraction, all the other results cannot improve upon the automatically created sub-document trained on document annotations.

Document Training		
	Binary	**3-Point**
Full Doc	69.5447	54.454
Optimal Text Extraction	75.0691	59.4621
Paragraph Training - Text Extraction Testing		
Word-based text extraction (n=20)	73.9379	55.112
Word-based text extraction (n=30)	72.7525	55.6334
Word-based text extraction (n=40)	73.0488	55.7445
Word-based text extraction (n=50)	72.9039	54.7676
Paragraph-based text extraction (n=0)	**75.2529**	55.1095
Paragraph Training - Aggregation Testing		
Sum of Predictions	72.1704	48.8426
Majority Voting	72.4105	46.743

Table 3. Binary and 3-Point document classification results using the paragraph annotations to train the classifier and segmenting the full document using word and paragraph windows during testing. The value of 'n' indicates the size of the text extraction window used.

As discussed in Section 4, an alternative approach is to train and test at the paragraph-level, but in a final step to aggregate the classification predictions to a single document score. The heading "Paragraph Training - Aggregation Testing" in Table 3 presents the results of using the paragraph data to do both training and testing with an aggregation step used to combine the predictions for multiple paragraphs to one overall document score. Although these approaches outperform document-based training approaches using full document representation, again they cannot match the performance of text extraction approaches trained using document annotations.

6 Conclusions

We have presented approaches which allow the classification of sentiment polarity at the document-level with the use of paragraph-level annotations. As these paragraph annotations relate to a more specific area of the document than the annotations at the document level it seems intuitive that they should be useful in providing more accurate information which can be leveraged by a machine

learning module. In particular the paragraph-level data would have a distinct advantage over document-level data when dealing with problems such as topic shift – a problem which we have shown to be present in the corpus that we use, based on an analysis in [11].

With the use of aggregation techniques, as well as with the integration of topic-based text extraction we were able to gain improvements over the standard approach of using the full document at both training and testing. However, when we compare these results to our previous work [11], which concentrated on segmenting the full document to create a more concise topic-based sub-document the results no longer provide any meaningful improvement – and for 3-point classification the results are significantly degraded.

When we consider the additional effort that is required in both generating these paragraph-level annotations, as well as the incorporation of additional strategies that are necessary to provide classification at the document-level (from the paragraph-level data), we would advise that the incorporation of this paragraph-level data is not necessarily beneficial. Of course, without the use of the additional topic-based text extraction that we performed, the paragraph-level data does produce an increase in performance. However, as the text extraction approach that we have previously proposed can be done fully automatically and only requires annotation to be done at the document-level we would have to advocate for its use and we see this as a further justification for its use in topic-based sentiment analysis.

Anecdotally, during the annotation process we observed a significant inconsistency in writing style in the blogs, even among respected bloggers. Some writers prefer one or two sentence paragraphs, while some prefer paragraphs of hundreds of words. We suspect that perhaps the irregularity with which paragraphs are used in user-generated content are used leads to a fundamental obstacle in interpreting them consistently in an automatic system.

Acknowledgments. This work is supported by Science Foundation Ireland under grant 07/CE/I1147, and by Enterprise Ireland under grant IP/2008/0549.

References

1. K. Ahmad, D. Cheng, and Y. Almas. Multi-lingual sentiment analysis of financial news streams. In *Proceedings of the 1st International Conference on Grid in Finance*, Palermo, 2006.
2. S. Das and M. Chen. Yahoo! for amazon: Sentiment extraction from small talk on the web. *Management Science*, 53(9):1375–1388, 2007.
3. A. Devitt and K. Ahmad. Sentiment polarity identification in financial news: A cohesion-based approach. In *Annual Meeting of the Association of Computational Linguistics*, page 984, 2007.
4. G. Forman. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3(1):1533–7928, 2003.
5. A. Funk, Y. Li, H. Saggion, K. Bontcheva, and C. Leibold. Opinion analysis for business intelligence applications. In *Proceedings of the first international workshop on Ontology-supported business intelligence*. ACM New York, NY, USA, 2008.

6. M. Koppel and I. Shtrimberg. Good news or bad news? let the market decide. *AAAI Spring Symposium on Exploring Attitude and* , Jan 2004.
7. M. Koppel and I. Shtrimberg. Good news or bad news?: Let the market decide. In *AAAI Spring Symposium on Exploring Attitude and Affect in Text*, pages 86–88. Springer, 2004.
8. K. Lerman, A. Gilder, M. Dredze, and F. Pereira. Reading the markets: Forecasting public opinion of political candidates by news analysis. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 473–480, Manchester, UK, August 2008. Coling 2008 Organizing Committee.
9. D. Lewis, Y. Yang, T. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.
10. C. Macdonald and I. Ounis. The trec blogs06 collection : Creating and analysing a blog test collection. Technical report, University of Glasgow, Department of Computing Science, 2006.
11. N. O'Hare, M. Davy, A. Bermingham, P. Ferguson, P. Sheridan, C. Gurrin, and A. F. Smeaton. Topic-dependent sentiment analysis of financial blogs. In *TSA'09 - 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion Measurement*, Nov 2009.
12. I. Ounis, C. Macdonald, and I. Soboroff. Overview of the trec-2008 blog track. 2008.
13. B. Pang and L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 115–124. Association for Computational Linguistics Morristown, NJ, USA, 2005.
14. B. Pang and L. Lee. *Opinion mining and sentiment analysis*. Now Publishers, 2008.
15. B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86, 2002.
16. M. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.
17. R. L. Santos, B. He, C. Macdonald, and I. Ounis. Integrating proximity to subjective sentences for blog opinion retrieval. In *ECIR '09: Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, pages 325–336, Berlin, Heidelberg, 2009. Springer-Verlag.
18. Y. Seki, D. K. Evans, L. Ku, L. Sun, H. Chen, and N. Kando. Overview of multilingual opinion analysis task at NTCIR-7. 2008.
19. P. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.
20. T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan. Opinionfinder: a system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, pages 34–35, Morristown, NJ, USA, 2005. Association for Computational Linguistics.
21. T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354. Association for Computational Linguistics Morristown, NJ, USA, 2005.